文章编号：2096-1472(2022)-02-22-03
DOI:10.19644/j.cnki.issn2096-1472.2022.002.006

基于影像组学和机器学习的脑部胶质瘤分级模型研究

王俊秀

(太原工业学院，山西 太原 030008)
✉wangjx@tit.edu.cn

摘　要：本文将影像组学的方法和机器学习算法结合起来，对脑部胶质瘤进行分级预测。利用BraTS2019公开数据集，从多模态MRI图像中分别提取肿瘤的448 维影像组学特征：肿瘤形态学特征、一阶灰度特征、纹理特征等；然后通过最小绝对收缩和选择算子(Lasso)算法筛选出15 个最佳的影像组学特征；最后根据筛选出的最佳特征集，利用随机森林分类算法构建脑部胶质瘤的分级预测模型。基于机器学习建立的模型在训练组患者中预测胶质瘤级别的准确率达到95.6%，ROC曲线下面积(AUC)达到0.99；在验证组患者中预测胶质瘤级别的准确率达到89.3%，AUC达到0.96。可见，基于机器学习算法，利用影像组学的方法可以对脑部肿瘤的高低级别进行准确的预测和分类。

关键词：肿瘤分级；影像组学；机器学习；随机森林
中图分类号：TP39　**文献标识码**：A

Research on Grading Model for Brain Glioma based on Radiomics and Machine Learning

WANG Junxiu

(*Taiyuan Institute of Technology*, *Taiyuan* 030008, *China*)
✉wangjx@tit.edu.cn

Abstract: This paper proposes to combine radiomics and machine learning algorithm to classify and predict the brain glioma. Based on BraTS2019 public dataset, 448-dimensional radiomics features of tumors are extracted from multimodal MRI (Magnetic Resonance Imaging) images, including tumor morphological features, first-order grayscale features, and texture features, etc. Then 15 best radiomics features are screened through the least absolute shrinkage and selection operator (Lasso) algorithm. Finally, according to the best screened feature set, the random forest classification algorithm is used to construct the brain glioma grading prediction Model. The accuracy of machine learning-based model is 95.6% and the area under the ROC (AUC) is 0.99 in the training group, and 89.3% and 0.96 in the validation group, respectively. Application of machine learning algorithm and radiomics realizes accurate prediction and classification of brain glioma level.

Keywords: brain glioma grading; radiomics; machine learning; random forest

1 引言(Introduction)

脑胶质瘤是大脑内部最常见的恶性肿瘤[1]，按照世界卫生组织的认定标准，根据胶质瘤的严重和恶性程度可划分为低级别胶质瘤(Low Grade Glioma，LGG)和高级别胶质瘤(High Grade Glioma，HGG)[2]。低级别胶质瘤为分化良好的胶质瘤，预后效果比较好。高级别胶质瘤为低分化胶质瘤，这类肿瘤为恶性肿瘤，患者预后效果不佳。胶质瘤的准确分级对患者的诊断、治疗方案的设计及预后非常重要[3]。影像组学研究是一个计算机和医学交叉研究的技术信息领域，它是指从各种类型的医学图像如CT、MRI、PET中提取高通量的数据信息，然后进一步地挖掘、分析和预测，最终可以帮助医生做出最准确的诊断与治疗[4]。影像组学包括获取图像、肿瘤区域分割、影像组学特征提取和分类预测模型构建等步骤。利用机器学习方法实现的影像组学已经很大程度上提高了医学诊断鉴别及预后预测的准确性[5]。

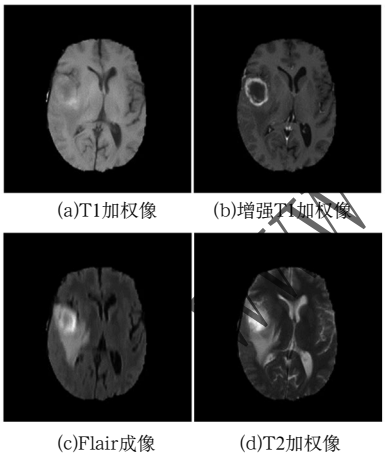
本文主要采用影像组学的方法和机器学习算法来解决脑部胶质瘤分级预测的问题。本研究使用了BraTS2019数据集中胶质瘤患者的术前MRI影像，采用影像组学方法提取影像

学特征；然后采用最小绝对收缩和选择算子(Least absolute shrinkage and selection operator，Lasso)对高维特征进行降维，筛选出最佳的影像学特征集；最后根据所选出的最佳特征集，通过随机森林(Random Forest，RF)算法建立胶质瘤高低级别分类模型。用受试者工作特征曲线(Receiver Operating Characteristic Curve，ROC曲线)来评价分类器模型的预测效果。

2 数据(Data)

磁共振成像(Magnetic Resonance Imaging，MRI)是大脑疾病诊断和治疗过程中的常规检查方法，在软组织检查中具有敏感性和卓越的图像对比度[6]。常见的头部MRI影像均包含T1加权成像、增强T1加权成像(T1ce)和T2加权成像，以及液体衰减反转恢复(Flair)成像等序列[7]。每个成像序列从不同的方面对肿瘤病灶进行描述，为脑胶质瘤诊断研究提供多个互补信息。

本文使用了BraTS2019数据集MRI影像进行研究，该数据集是2019 年脑部肿瘤分割竞赛数据集(Brain Tumor Segmentation Challenge 2019，BraTS2019)[8]，包括76 例低级别胶质瘤MRI影像和259 例高级别胶质瘤MRI影像。数据集中包含T1加权像、增强T1加权像、T2加权像和液体衰减反转恢复序列像四个模态的MRI影像，另外每例病人还包括医学专家手工标记的肿瘤区域和肿瘤分级的情况。所有的影像数据都进行了图像预处理，包括配准、图像插值和重采样等。图1为BraTS2019数据集中一例患者的脑部MRI影像。



(a)T1加权像　(b)增强T1加权像

(c)Flair成像　(d)T2加权像

图1 BraTS2019数据集中脑部MRI影像

Fig.1 Brain MRI image in BraTS2019 dataset

数据集中每例患者的肿瘤区域都是由多个经验丰富的医生按照相同的标注规范进行分割和验证的。肿瘤区域按照病理一般分为四个区域：(1)增强肿瘤核心区(Enhanced Core)；(2)肿瘤周围水肿区(Edema)；(3)非增强肿瘤核心区(Non-enhancing Solid Core)；(4)坏死区/囊性核心区(Necrotic/Cystic Core)。其中(3)和(4)为真实的胶质瘤组织，合并为一个区域，简称为NET区域；增强的肿瘤核心区域简称为ET区域；肿瘤周围水肿区域简称为ED区域。图2为一例患者病灶区域分割的图像，其中浅灰色ED区域，深灰色为NET区域，白色为ET区域。

图2 病灶区域分割的图像

Fig.2 Image of the lesion segmentation

3 方法(Methods)

3.1 特征提取

本文根据近几年研究人员提出的对脑部胶质瘤图像提取的特征，从四种模态图像(T1、T2、T1ce和Flair)的不同病灶区域中分别提取了肿瘤形态学特征、一阶特征及纹理特征，共计448 个影像组学特征，每种类型的特征从不同的方面对图像进行描述，解析了图像的隐含特点。(1)肿瘤形态学特征：提取肿瘤原始空间的三维特征，可以量化肿瘤的形状和大小。(2)一阶灰度特征：由感兴趣区域影像特征值直方图计算而来，可以定量地描述图像的信号强度分布。(3)纹理特征：可以对脑胶质瘤内异质性进行定量刻画。纹理特征包括：灰度共生矩阵纹理特征(GLCM)[9]、灰度相关矩阵纹理特征(GLDM)[10]、灰度游程矩阵纹理特征(GLRLM)[11]、灰度区域大小矩阵纹理特征(GLSZM)[12]、邻域灰度差矩阵纹理特征(NGTDM)[13]。

3.2 特征选择

本文主要通过采用$L1$正则化Lasso回归分析模型来进行最佳特征筛选[14]。Lasso是一种用于变量压缩和估计的方法，它可以有效地将高维变量降到十几维甚至更少，同时不影响模型的预测能力。目前Lasso算法已经被广泛应用在高维数据的降维和回归分析中，特别是影像组学的特征工程领域。本文采用Lasso回归模型选择出和胶质瘤高低级别最相关的影像组学特征。简单线性回归模型的定义如下：

$$f(x)=\sum_{j=1}^{p}w^{j}x^{j} \tag{1}$$

式(1)中，x表示样本，w表示要拟合的参数，p表示样本特征的维数。应用二次损失来表示目标函数，即：

$$J(w)=\frac{1}{n}\sum_{i=1}^{n}(y_i-f(x_i))^2=\frac{1}{n}\|y-Xw\|^2 \tag{2}$$

式(2)中，X是特征矩阵，$X=(x_1,\cdots,x_n)^T\in R^{n\times p}$；$y$是由标签构成的列向量，$y=(y_1,\cdots,y_n)^T$。式(2)的解析解为：

$$\hat{w}=(X^TX)^{-1}Xy \tag{3}$$

然而，若$p>>n$，即特征的维数远远超出了样本的个数，矩阵X^TX就不能完全满秩，此时就没有解。通过Lasso正则化，得到一个目标函数：

$$J_L(w)=\frac{1}{n}\|y-Xw\|^2+\lambda\|w\|_1 \tag{4}$$

式(4)等价为：

$$\min_{w}\frac{1}{n}\|y-Xw\|^2,\quad \text{s.t.}\ \|w\|_1\le C \tag{5}$$

为了去掉冗余的特征，本文采用$L1$正则方法进行特征压缩。当特征维度是2时，我们可以直接在平面上绘制出目标函数的等高线，而取值范围是平面上半径为C的$L1$范数圆，等高线与$L1$范数圆的交点就是最优解。而更高维的情况下，等

高线与$L1$范数球的交点除了角点之外还可能在很多边的轮廓线上，同样具有稀疏性。

3.3 分级预测模型的构建

基于筛选出的影像组学特征和随机森林构建模型。随机森林算法是一种通过采用Bagging的算法将多个无关联的决策树组合在一起，以投票机制进行分类的有监督学习算法。随机森林算法的泛化能力强，分类性能比较好，广泛应用于各种分类任务中[15]。

决策树是一种树形结构的分类器。在构建决策树时，树中的每个节点都要选择最优的特征对当前样本进行分类，直到决策树能够满足所需要的建树停止的条件。当把一个样本输入决策树中时，可以自动确定一条从根节点开始到叶节点的唯一路径，最后叶节点也就是这个样本的类别[16]。随机森林中构建的每一棵决策树都可以是一个分类器，当把一个样本输入随机森林中时，M棵决策树会得到M个分类结果，根据所有决策树的分类结果，把次数最高的类别作为最终分类结果。本文中随机森林算法按照脑部胶质瘤高低级别的分布情况进行随机抽样。模型训练的过程中可采用并行方法，这样使得模型训练速度快。

决策树的深度直接影响随机森林分类器模型的性能，如果决策树的深度过大会导致分类模型过拟合，而决策树的深度过小又会导致分类模型欠拟合。决策树的数量也会影响随机森林的分类准确率。在实现过程中，采用TPOT(Tree-based Pipeline Optimization Tool)框架实现随机森林的自动机器学习，以确定最优的决策树个数和决策树深度。TPOT框架是由美国宾夕法尼亚大学自主研究和设计开发的一个自动机器学习的技术框架[17]。它是一种基于遗传算法的Python自动机器学习工具。TPOT能够进行自动算法选择、自动参数优化，为当前数据集找到最优的算法及其参数。

4 结果(Results)

将BraTS2019数据集的335 例胶质瘤患者随机分为训练集(75%)和验证集(25%)。每个MRI模态提取112 个特征，包括19 个肿瘤形态学特征、18 个一阶灰度特征及75 个纹理特征，四个模态共提取448 个影像组学特征。448 个影像组学特征的Lasso系数分布如图3所示。

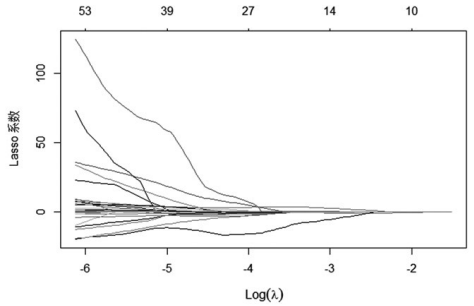


图3 448 个影像组学特征的Lasso系数分布

Fig.3 Lasso coefficient distribution of 448 radiomics features

使用Lasso回归模型对448 个影像组学特征进行压缩，通过交叉验证和二项式偏差最小化确定Lasso回归模型中惩罚系数λ的最优值，如图4所示。同时筛选出系数非零的最佳特征变量，如表1所示，共筛选出15 个最佳影像组学特征。

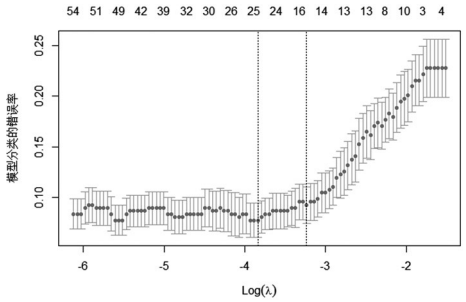


图4 Lasso回归模型交叉验证的结果图

Fig.4 Cross-validation results of Lasso regression model

表1 筛选出的15 个最佳影像组学特征

Tab.1 15 best selected radiomics features

模态	特征	数量
T1	t1-diagnostics_Mask-original_VolumeNum t1-original_shape_MeshVolume t1-original_shape_Sphericity t1-original_glrlm_RunEntropy t1-original_glrlm_RunLengthNonUniformity t1-original_glszm_LargeAreaLowGrayLevelEmphasis	6
T1ce	t1ce-diagnostics_Mask-original_VolumeNum t1ce-original_shape_Sphericity t1ce-original_glcm_Idn t1ce-original_glcm_Imc1 t1ce-original_gldm_DependenceNonUniformity	5
T2	t2-diagnostics_Mask-original_VolumeNum t2-original_shape_MeshVolume t2-original_shape_Sphericity	3
Flair	flair-original_firstorder_Skewness	1

本文采用TPOT框架实现随机森林分类模型的自动机器学习，从而确定最优的随机森林分类器参数：决策树的最大深度(max_depth)为9，基学习器的个数(n_estimators)为100。基于15 个最优的影像组学特征，通过TPOT构建的随机森林分类器来预测胶质瘤高低级别，在训练组患者中预测胶质瘤级别的准确率达到95.6%，在验证组患者中预测胶质瘤级别的准确率达到89.3%。绘制ROC曲线来评价分级模型，训练组的曲线下面积AUC为0.99，验证组的AUC为0.96。ROC曲线如图5所示。

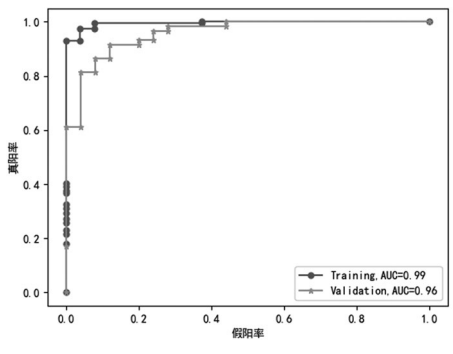


图5 模型在训练组和验证组的受试者工作特征曲线(ROC)

Fig.5 Subjects operating characteristic curves of the model in training group and validation group

5 结论(Conclusion)

本文采用影像组学的方法和机器学习算法对脑部胶质瘤进行高低级别分级预测。从MRI影像的四个模态上提取了一

系列胶质瘤的影像组学特征，使用Lasso回归模型进行筛选，得到和胶质瘤级别密切相关的最佳影像组学特征集，并基于所选的特征建立了随机森林分类器的预测模型。我们发现，该模型在训练组和验证组中均实现了胶质瘤高低级别的有效预测。

参考文献(References)

[1] MORGAN L L. The epidemiology of glioma in adults: a "state of the science" review[J]. Neuro-Oncology, 2015, 17(4): 623-624.

[2] LOUIS D N, PERRY A, REIFENBERGER G, et al. The 2016 world health organization classification of tumors of the central nervous system: A summary[J]. Acta Neuropathologica, 2016, 131(6):803-820.

[3] JANG K, RUSSO C, IEVA A D. Radiomics in gliomas: Clinical implications of computational modeling and fractal-based analysis[J]. Neuroradiology, 2020, 62(7):771-790.

[4] LAMBIN P, RIOS-VELAZQUEZ E, LEIJENAAR R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis[J]. European Journal of Cancer, 2012, 48(4):441-446.

[5] YIP S S, AERTS H J. Applications and limitations of radiomics[J]. Physics in Medicine and Biology, 2016, 61(13):150-166.

[6] FOUKE S J, BENZINGER T, GIBSON D, et al. The role of imaging in the management of adults with diffuse low grade glioma: A systematic review and evidence-based clinical practice guideline[J]. Journal of Neuro Oncology, 2015, 125(3):457-479.

[7] 贾颖,杜学松,陈君辉,等.基于常规MRI的定量影像学特征用于胶质瘤分级诊断[J].中国医学影像技术,2018,034(008): 1137-1142.

[8] MENZE B H, JAKAB A, BAUER S, et al. The multimodal brain tumor image segmentation benchmark (BRATS)[J]. IEEE Transactions on Medical Imaging, 2015, 34(10):1993-2024.

[9] HARALICK R M, SHANMUGAM K, DINSTEIN I. Textural features for image classification[J]. Studies in Media and Communication, 1973, 3(6):610-621.

[10] SUN C, WEE W G. Neighboring gray level dependence matrix for texture classification[J]. Computer Vision Graphics and Image Processing, 1983, 23(3):341-352.

[11] GALLOWAY M. Texture analysis using gray level run lengths[J]. Computer Graphics and Image Processing, 1975, 4(2):172-179.

[12] THIBAULT G, ANGULO J, MEYER F. Advanced statistical matrices for texture characterization: Application to cell classification[J]. IEEE Transactions on Biomedical Engineering, 2014, 61(3):630-637.

[13] AMADASUN M, KING R. Textural features corresponding to textural properties[J]. IEEE Transactions Systems, Man, and Cybernetics, 1989, 19(5):1264-1274.

[14] TIBSHIRANI R. Regression shrinkage and selection via the lasso: A retrospective[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2011, 73(3): 267-288.

[15] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1):5-32.

[16] 王奕森,夏树涛.集成学习之随机森林算法综述[J].信息通信技术,2018(1):51-57.

[17] OLSON R S, MOORE J H. TPOT: A tree-based pipeline optimization tool for automating machine learning[C]// HUTTER F, KOTTHOFF L, VANSCHOREN J. Automatic Machine Learning. Cham: Springer, 2016, 64:66-74.

作者简介：

王俊秀(1987-)，女，博士生，讲师.研究领域：图像处理，人工智能，医学图像.

(上接第28页)

参考文献(References)

[1] 李豪,彭庆,谭美容.面向乘客策略行为的航空公司舱位控制与动态定价模型[J].控制与决策,2018,33(07):1295-1302.

[2] 秦瑛,霍佳震,陈军,等.基于需求转移的航空公司座位分配博弈模型[J].统计与决策,2016,32(02):56-60.

[3] 闫振英,韩宝明,李晓娟,等.考虑旅客选择行为的高铁席位动态控制策略[J].交通运输系统工程与信息,2019,19(01): 118-124.

[4] 衡红军,黄小荣,王治宝.EMSR在航空收益管理系统中的应用[J].计算机工程,2003(12):139-141.

[5] 樊玮,苏秋波.基于分布估计算法的多航段座位分配模型[J].信息与控制,2012,41(06):774-778,785.

[6] PETER P B. Application of a probabilistic decision model to airline seat inventory control[J]. Operation Research, 1989, 37(02):183-197.

[7] GUILLERMO G, HUSEYIN T. Revenue management and pricing analytics[M]. New York: Springer, 2019:23-47.

[8] 张正,贾小林.面向NB-IOT智能设备动态链接库的远程技术研究及应用[J].计算机应用与软件,2021,38(06):170-175.

[9] 刘琴.大数据分析下分布式数据流处理技术研究[J].软件工程,2019,22(12):44-46.

作者简介：

王洪建(1966-)，男，硕士，高级工程师.研究领域：计算机应用，智能优化.